Model averaging and dimension selection for the singular value decomposition

Peter D. Hoff *
Technical Report no. 494
Department of Statistics
University of Washington

January 10, 2006

Abstract

Many multivariate data analysis techniques for an $m \times n$ matrix $\mathbf{Y}$ are related to the model $\mathbf{Y} = \mathbf{M} + \mathbf{E}$, where $\mathbf{Y}$ is an $m \times n$ matrix of full rank and $\mathbf{M}$ is an unobserved mean matrix of rank $K < (m \wedge n)$. Typically the rank of $\mathbf{M}$ is estimated in a heuristic way and then the least-squares estimate of $\mathbf{M}$ is obtained via the singular value decomposition of $\mathbf{Y}$, yielding an estimate that can have a very high variance. In this paper we suggest a model-based alternative to the above approach by providing prior distributions and posterior estimation for the rank of $\mathbf{M}$ and the components of its singular value decomposition.

Some key words: Carlson's hypergeometric function, directional data, factor analysis, interaction, model selection, relational data, social network, Steifel manifold.

1 Introduction

Every $m \times n$ matrix $\mathbf{M}$ has a representation of the form $\mathbf{M} = \mathbf{UDV}'$ where, in the case $m \geq n$,

- $\mathbf{U}$ is an $m \times n$ matrix with orthonormal columns;
- $\mathbf{V}$ is an $n \times n$ matrix with orthonormal columns;
- $\mathbf{D}$ is an $n \times n$ diagonal matrix, with diagonal elements $\{d_1, \ldots, d_n\}$ typically taken to be a decreasing sequence of non-negative numbers.

*Departments of Statistics, Biostatistics and the Center for Statistics and the Social Sciences, University of Washington, Seattle, Washington 98195-4322, U.S.A.. Email: hoff@stat.washington.edu. This research was supported by Office of Naval Research grant N00014-02-1-1011 and National Science Foundation grant SES-0417559. The author thanks David Hoff for helpful discussions.

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
10 JAN 2006		**00-01-2006 to 00-01-2006**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Model averaging and dimension selection for the singular value decomposition	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **University of Washington,Department of Statistics,Box 354322,Seattle,WA,98195-4322**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**		**24**	

The triple $\{\mathbf{U}, \mathbf{D}, \mathbf{V}\}$ is called the singular value decomposition of $\mathbf{M}$. The squared elements of the diagonal of $\mathbf{D}$ are the eigenvalues of $\mathbf{M}'\mathbf{M}$ and the columns of $\mathbf{V}$ are the corresponding eigenvectors. The matrix $\mathbf{U}$ can be obtained from the first n eigenvectors of $\mathbf{M}\mathbf{M}'$. The number of non-zero elements of $\mathbf{D}$ is the rank of $\mathbf{M}$.

Many data analysis procedures for matrix-valued data $\mathbf{Y}$ are related to the singular value decomposition, partly due to its appealing interpretation as a multiplicative model based on row and column factors. Given a model of the form $\mathbf{Y} = \mathbf{M} + \mathbf{E}$, the elements of $\mathbf{Y}$ can be written $y_{i,j} = \mathbf{u}_i'\mathbf{D}\mathbf{v}_j + e_{i,j}$, where $\mathbf{u}_i$ and $\mathbf{v}_j$ are the ith and jth rows of $\mathbf{U}$ and $\mathbf{V}$ respectively. Models of this type play a role in the analysis of relational data (Harshman et al., 1982), biplots (Gabriel 1971, Gower and Hand 1996) and in reduced-rank interaction models for factorial designs (Gabriel 1978, 1998). The model is also closely related to factor analysis, where the row vectors of $\mathbf{Y}$ are modeled as i.i.d. samples from the model $\mathbf{y}_i = \mathbf{u}_i\mathbf{D}\mathbf{V}' + \mathbf{e}_i$. In this situation, the goal is typically to represent the covariance across the n columns by their relationship to $K < n$ unobserved latent factors.

The singular value decomposition also plays a role in parameter estimation for the above model: Assuming the rank of the mean matrix $\mathbf{M}$ is $K < n$ and letting $(\hat{\mathbf{U}}, \hat{\mathbf{D}}, \hat{\mathbf{V}})$ be the singular value decomposition of the data matrix $\mathbf{Y}$, the least-squares estimate of $\mathbf{M}$ (and maximum likelihood estimate under Gaussian noise) is given by $\hat{\mathbf{M}}_K = \hat{\mathbf{U}}_{[,1:K]}\hat{\mathbf{D}}_{[1:K,1:K]}\hat{\mathbf{V}}'_{[,1:K]}$, where $\hat{\mathbf{U}}_{[,1:K]}$ denotes the first K columns of $\hat{\mathbf{U}}$ and $\hat{\mathbf{D}}_{[1:K,1:K]}$ denotes the first K rows and columns of $\hat{\mathbf{D}}$ (Householder and Young 1938, Gabriel 1978). In applications such as signal processing, image analysis, and more recently large-scale gene expression data, representing a noisy data matrix $\mathbf{Y}$ by $\hat{\mathbf{M}}_K$ with $K << n$ has the effect of capturing the main patterns of $\mathbf{Y}$ while eliminating much of the noise.

Despite its utility and simplicity, two issues limit the use of the singular value decomposition as an estimation procedure. The first is that the rank K of the approximating mean matrix $\hat{\mathbf{M}}_K$ must be specified. Standard practice is to plot the singular values of $\mathbf{Y}$ in decreasing order and then select K to be the index where the last "large gap" occurs. The second issue is that, even if the rank is chosen correctly, the least-squares estimate has a very high variance: The value of $||\hat{\mathbf{M}}_K||^2$ is equal to the sum of the first K eigenvalues of $\mathbf{Y}'\mathbf{Y}$, which has expectation $E(\mathbf{Y}'\mathbf{Y}) = \mathbf{M}'\mathbf{M} + m\sigma^2\mathbf{I}$ (where σ^2 is the variance of the elements of $\mathbf{E}$). As a result, the entries of $\hat{\mathbf{M}}_K$ can be much larger in magnitude than the corresponding entries in $\mathbf{M}$.

Philosophical debates aside, Bayesian methods often provide sensible procedures for model selection and high-dimensional parameter estimation. For the model described above, a Bayesian procedure would provide a mapping from a prior distribution $p(\mathbf{U}, \mathbf{D}, \mathbf{V}, \sigma^2)$ to a posterior distribution $p(\mathbf{U}, \mathbf{D}, \mathbf{V}, \sigma^2|\mathbf{Y})$. Of primary interest might be functions of this posterior distribution, such as $E(\mathbf{M}|\mathbf{Y})$ or the marginal posterior distribution of the rank $p(K|\mathbf{Y}) \propto p(K)p(\mathbf{Y}|K)$. Both of these quantities require integration over the complicated, high-dimensional space of $\{\mathbf{U}, \mathbf{D}, \mathbf{V}\}$

for each value of K. In the related factor analysis model where the elements of $\mathbf{U}$ are modeled as independent normal random variables, the difficulty in calculating marginal probabilities has led to the development of approximate Bayesian procedures: Rajan and Rayner (1997) provide a coarse approximation to the marginal probability $p(\mathbf{Y}|K)$ by plugging in maximum-likelihood estimates. Minka (2000) improves on this by providing a Laplace approximation to the desired marginal probability. Both of these procedures rely on asymptotic approximations, and do not provide Bayesian estimates of $\mathbf{M}$ once the dimension has been selected. In contrast, Lopes and West (2004) provide a unified procedure that provides both model selection and parameter estimation for the factor analysis model, although their approach to model selection requires a complicated two-stage reversible-jump MCMC algorithm: The first stage runs separate MCMC algorithms for each rank K to be considered, and the second stage runs a Markov chain between ranks, using results of the first stage to approximate marginal probabilities.

In many situations the row heterogeneity and column heterogeneity of $\mathbf{Y}$ are of equal interest. In these cases, the factor analysis approaches mentioned above may be less appropriate than a model for the singular value decomposition of $\mathbf{M}$. The goal of this paper is to provide a method of estimation and inference for such a model. Specifically, this paper provides the necessary calculations for Bayesian estimation and model averaging for a mean matrix $\mathbf{M}$ by way of its singular value decomposition $\{\mathbf{U}, \mathbf{D}, \mathbf{V}\}$. In Section 2 we discuss prior distributions for $\{\mathbf{U}, \mathbf{D}, \mathbf{V}\}$ given a fixed rank K, and show how the uniform distribution for $\mathbf{U}$ (the invariant measure on the Steifel manifold) may be specified in terms of the full conditional distributions of its column vectors. Section 3 presents a Gibbs sampling scheme for parameter estimation when the rank of $\mathbf{M}$ is specified. In the case of unspecified rank, estimation can be achieved via a prior distribution which allows the diagonal elements of $\mathbf{D}$ to each be zero with non-zero probability. A Markov chain Monte Carlo algorithm which moves between models with different ranks is constructed via a Gibbs sampling scheme which samples each singular value d_j from its conditional distribution. This is done marginally over $\mathbf{U}_{[,j]}$ and $\mathbf{V}_{[,j]}$, and requires a complicated but manageable integration. Section 5 presents a small simulation study that examines the sampling properties of the Bayesian procedure. It is shown that the procedure is able to estimate the true rank of $\mathbf{M}$ reasonably well for a variety of matrix sizes, and the squared error of the Bayes estimate $E(\mathbf{M}|\mathbf{Y})$ is typically much lower than that of the least squares estimator. Model extensions for non-normal data are described in Section 6, along with an example analysis of binary relational data. A discussion follows in Section 7.

2 The SVD model and prior distributions

As described above, our model for an $m \times n$ data matrix is $\mathbf{Y} = \mathbf{M} + \mathbf{E}$, where $\mathbf{M}$ is a rank K matrix and $\mathbf{E}$ is a matrix of i.i.d. mean-zero normally-distributed noise. We induce a prior distribution on

the matrix $\mathbf{M}$ by way of a prior distribution on the components of its singular value decomposition $\{\mathbf{U}, \mathbf{D}, \mathbf{V}\}$.

For a given rank K, we can take $\mathbf{U}$ to be an $m \times K$ orthonormal matrix. The set of such matrices is called the Steifel manifold and is denoted $\mathcal{V}_{K,m}$. A natural, non-informative prior distribution for $\mathbf{U}$ is the uniform distribution on $\mathcal{V}_{K,m}$, which is the unique probability measure on $\mathcal{V}_{K,m}$ that is invariant under left and right orthogonal transformations. As discussed in Chikuse (2003, Section 2.5), a sample $\mathbf{U}$ from the uniform distribution on the Steifel manifold $\mathcal{V}_{K,m}$ may be obtained by first sampling an $m \times K$ matrix $\mathbf{X}$ of independent standard normal random variables and then setting $\mathbf{U} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1/2}$. Although this construction is straightforward, it doesn't explicitly specify conditional distributions of the form $p(\mathbf{U}_{[,j]}|\mathbf{U}_{[,-j]})$, which are quantities that will be required for the estimation procedure outlined in Section 3. We now derive these conditional distributions via a new iterative method of generating samples from the uniform distribution on $\mathcal{V}_{K,m}$.

Let $\mathbf{U}_{[,A]}$ denote the columns of $\mathbf{U}$ corresponding to a subset of column labels $A \subset \{1, \ldots, K\}$, and let $\mathbf{N}_A$ be any $m \times (m - |A|)$ matrix whose columns form an orthonormal basis for the null space of $\mathbf{U}_{[,A]}$. A random $\mathbf{U} \in \mathcal{V}_{K,m}$ can be constructed as follows:

1. Sample $\mathbf{u}_1$ uniformly from the m-sphere and set $\mathbf{U}_{[,1]} = \mathbf{u}_1$;

2. Sample $\mathbf{u}_2$ uniformly from the $(m-1)$-sphere and set $\mathbf{U}_{[,2]} = \mathbf{N}_{\{1\}}\mathbf{u}_2$;

$\vdots$

K. Sample $\mathbf{u}_K$ uniformly from the $(m - K + 1)$-sphere and set $\mathbf{U}_{[,K]} = \mathbf{N}_{\{1,\ldots,K-1\}}\mathbf{u}_K$.

By construction this procedure generates an $m \times K$ matrix $\mathbf{U}$ having orthonormal columns. The following result also holds:

Proposition 1 *The probability distribution of* $\mathbf{U}$ *is the uniform probability measure on* $\mathcal{V}_{K,m}$.

A proof is provided in the Appendix. Since this probability distribution is invariant under left and right orthogonal transformations of $\mathbf{U}$ (see, for example, Chikuse 2003), it follows that the rows and columns of $\mathbf{U}$ are exchangeable. As a result, the conditional distribution of $\mathbf{U}_{[,j]}$ given any subset A of columns of $\mathbf{U}$ is equal to the distribution of $\mathbf{N}_A\mathbf{u}_j$, where $\mathbf{u}_j$ is uniformly distributed on the $(m - |A|)$-sphere. This fact facilitates the Gibbs sampling of the columns of $\mathbf{U}$ and $\mathbf{V}$ from their full conditional distributions, as described in Section 3.

For a given rank K, the non-zero singular values $\{d_1, \ldots, d_K\}$ which make up the diagonal of $\mathbf{D}$ determine the magnitude of the mean matrix, in that $||\mathbf{M}||^2 = \sum_{k=1}^{K} d_k^2$. We model these non-zero values as being samples from a normal population with mean μ and precision (inverse-variance) ψ. Conjugate prior distributions for these parameters include a normal distribution with mean μ_0 and variance v_0^2 for μ, and a gamma$(\eta_0/2, \eta_0\tau_0^2/2)$ distribution for ψ, parameterized so

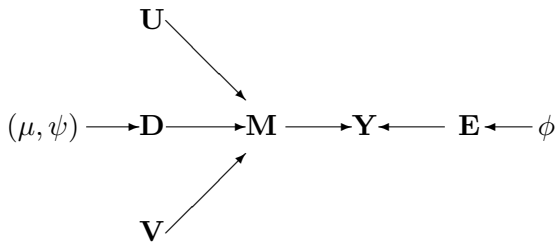

Figure 1: A graphical representation of the model

that ψ has expectation $1/\tau_0^2$. This parameterization of the singular values differs slightly from that of the usual singular value decomposition, in that the values $\{d_1, \ldots, d_K\}$ are not restricted to be non-negative here. A model enforcing this restriction is possible, but adds a small amount of computational difficulty without any modeling benefit (if $\mathbf{A}$ is a diagonal matrix of ± 1's, then $p(\mathbf{Y}|\mathbf{U}, \mathbf{D}, \mathbf{V}) = p(\mathbf{Y}|\mathbf{UA}, \mathbf{AD}, \mathbf{V})$). Finally, the elements of $\mathbf{E}$ are modeled as i.i.d. normal random variables with mean zero and variance $1/\phi$. The prior distribution for the precision ϕ is taken to be gamma$(\nu_0/2, \nu_0\sigma_0^2/2)$. A graphical representation of the model and parameters is given in Figure 1. Choices for hyperparameters $\{(\mu_0, v_0^2), (\eta_0, \tau_0^2), (\nu_0, \sigma_0^2)\}$ are discussed in Section 5.

3 Gibbs sampling for the fixed-rank model

A Markov chain with $p(\mathbf{U}, \mathbf{D}, \mathbf{V}, \phi, \mu, \psi|\mathbf{Y}, K)$ as its stationary distribution can be constructed via a Gibbs sampling procedure, which iteratively samples ϕ, μ, ψ and the columns of $\mathbf{U}$, $\mathbf{D}$ and $\mathbf{V}$ from their full conditional distributions. These samples can be used to approximate the joint posterior distribution and estimate posterior quantities of interest (see, for example, Tierney 1994).

The full conditional distributions for ϕ, μ, ψ and the elements of $\mathbf{D}$ are standard and are provided below without derivation. Less standard are the full conditional distributions of the columns of $\mathbf{U}$ and $\mathbf{V}$. To derive these, consider the form of $p(\mathbf{Y}|\mathbf{U}, \mathbf{D}, \mathbf{V}, \phi)$ as a function of $\mathbf{U}_{[,j]}, \mathbf{V}_{[,j]}$ and $d_j \equiv \mathbf{D}_{[j,j]}$. Letting $\mathbf{E}_{-j} = \mathbf{Y} - \mathbf{U}_{[,-j]}\mathbf{D}_{[-j,-j]}\mathbf{V}'_{[,-j]}$, we have

$$\begin{aligned} ||\mathbf{Y} - \mathbf{UDV}'||^2 &= ||\mathbf{E}_{-j} - d_j\mathbf{U}_{[,j]}\mathbf{V}'_{[,j]}||^2 \\ &= ||\mathbf{E}_{-j}||^2 - 2d_j\mathbf{U}'_{[,j]}\mathbf{E}_{-j}\mathbf{V}_{[,j]} + ||d_j\mathbf{U}_{[,j]}\mathbf{V}'_{[,j]}||^2 \\ &= ||\mathbf{E}_{-j}||^2 - 2d_j\mathbf{U}'_{[,j]}\mathbf{E}_{-j}\mathbf{V}_{[,j]} + d_j^2. \end{aligned}$$

It follows that $p(\mathbf{Y}|\mathbf{U}, \mathbf{D}, \mathbf{V}, \phi)$ can be written

$$p(\mathbf{Y}|\mathbf{U}, \mathbf{D}, \mathbf{V}, \phi) = \left(\frac{\phi}{2\pi}\right)^{mn/2} \exp\{-\frac{1}{2}\phi||\mathbf{E}_{-j}||^2 + \phi d_j\mathbf{U}'_{[,j]}\mathbf{E}_{-j}\mathbf{V}_{[,j]} - \frac{1}{2}\phi d_j^2\}. \tag{1}$$

Recall that conditional on $\mathbf{U}_{[,-j]}$, $\mathbf{U}_{[,j]} \stackrel{d}{=} \mathbf{N}^u_{\{-j\}}\mathbf{u}_j$, where $\mathbf{N}^u_{\{-j\}}$ is a basis for the null space of columns of $\mathbf{U}_{[,-j]}$ and $\mathbf{u}_j$ is uniform on the $m-(K-1)$-sphere. From (1), we see that the full conditional distribution of $\mathbf{u}_j$ is proportional to $\exp\{\mathbf{u}_j'\boldsymbol{\mu}\}$, where $\boldsymbol{\mu} = \phi d_j \mathbf{N}^{u\prime}_{\{-j\}}\mathbf{E}_{-j}\mathbf{V}_{[,j]}$. This is a von Mises-Fisher distribution on the $m-(K-1)$-sphere with parameter $\boldsymbol{\mu}$. A sample of $\mathbf{U}_{[,j]}$ from its full conditional distribution can therefore be generated by sampling $\mathbf{u}_j$ from the von Mises-Fisher distribution and then setting $\mathbf{U}_{[,j]} = \mathbf{N}^u_{\{-j\}}\mathbf{u}_j$. The full conditional distribution of $\mathbf{V}_{[,j]}$ is derived similarly. In general, the von Mises-Fisher distribution on the p-sphere with parameter $\boldsymbol{\mu} \in \mathbb{R}^p$ has density $c_p(||\boldsymbol{\mu}||)\exp\{\mathbf{u}'\boldsymbol{\mu}\}$ and is denoted vMF($\boldsymbol{\mu}$), and the uniform distribution on the sphere is denoted vMF($\mathbf{0}$). The normalizing constants for these two cases are

$$c_p(\kappa) = (2\pi)^{-p/2}\frac{\kappa^{p/2-1}}{I_{p/2-1}(\kappa)} \text{ for } \kappa > 0, \quad c_p(0) = \frac{\Gamma(p/2)}{2\pi^{p/2}} \text{ for } \kappa = 0,$$

where $I_\nu(x)$ is the modified Bessel function of the first kind. R-code for sampling from this distribution is provided at my website.

Summarizing these results, a Markov chain with the desired stationary distribution can be constructed by iterating the following procedure:

- For $j \in \{1, \ldots, K\}$,
  - sample $(\mathbf{U}_{[,j]}|\mathbf{Y}, \mathbf{U}_{[,-j]}, \mathbf{D}, \mathbf{V}, \phi) \stackrel{d}{=} \mathbf{N}^u_{\{-j\}}\mathbf{u}_j$, where $\mathbf{u}_j \sim$ vMF$(\phi d_j \mathbf{N}^{u\prime}_{-j}\mathbf{E}_{-j}\mathbf{V}_{[,j]})$;
  - sample $(\mathbf{V}_{[,j]}|\mathbf{Y}, \mathbf{U}, \mathbf{D}, \mathbf{V}_{[,-j]}, \phi) \stackrel{d}{=} \mathbf{N}^v_{\{-j\}}\mathbf{v}_j$, where $\mathbf{v}_j \sim$ vMF$(\phi d_j \mathbf{U}'_{[,j]}\mathbf{E}_{-j}\mathbf{N}^v_{\{-j\}})$;
  - sample $(d_j|\mathbf{Y}, \mathbf{U}, \mathbf{D}_{[-j,-j]}, \mathbf{V}, \phi, \mu, \psi) \sim$ normal $[(\mathbf{U}'_{[,j]}\mathbf{E}_{-j}\mathbf{V}_{[,j]}\phi + \mu\psi)/(\phi+\psi), 1/(\phi+\psi)]$;
- sample $(\phi|\mathbf{Y}, \mathbf{U}, \mathbf{D}, \mathbf{V}) \sim$ gamma$[(\nu_0 + mn)/2, (\nu_0\sigma_0^2 + ||\mathbf{Y} - \mathbf{UDV}'||^2)/2]$;
- sample $(\mu|\mathbf{D}, \psi) \sim$ normal$[(\psi\sum d_j + \mu_0/v_0^2)/(\psi K + 1/v_0^2), 1/(\psi K + 1/v_0^2)]$
- sample $(\psi|\mathbf{D}, \mu) \sim$ gamma$[(\eta_0 + K)/2, (\eta_0\tau_0^2 + \sum(d_j - \mu)^2)/2]$;

4 The variable-rank model

4.1 Prior distributions

In this section we extend the model of Section 2 to the case where the rank K is to be estimated. This requires comparisons between models with parameter spaces of different dimension. Two standard ways of viewing such problems are as follows:

- Conceptualize a different parameter space for each value of K, i.e., conditional on K, the mean matrix is $\mathbf{UDV}'$ where the dimensions of $\mathbf{U}, \mathbf{D}$ and $\mathbf{V}$ are $m \times K$, $K \times K$ and $n \times K$ respectively.

- Parameterize $\mathbf{U}, \mathbf{D}$ and $\mathbf{V}$ to be of dimensions $m \times n$, $n \times n$ and $n \times n$, but allow for columns of these matrices to be identically zero. In this parameterization, $K = \sum_{j=1}^{n} 1(d_j \neq 0)$.

Each of these two approaches has its own notational and conceptual hurdles, and which one to present is to some extent a matter of style (see Green 2003 for a discussion). Given a prior distribution on K, the first approach can be formulated by using the prior distributions of Section 2 as the conditional distributions of $\mathbf{U}, \mathbf{D}$ and $\mathbf{V}$ given K. The second approach can be made equivalent to the first as follows:

1. Let $\tilde{\mathbf{U}}, \tilde{\mathbf{D}}, \tilde{\mathbf{V}}$ have the prior distributions described in Section 2 with $\tilde{K} = n$;
2. Let $\{s_1, \ldots, s_n\} \sim p(K = \sum s_j) \times \binom{n}{\sum s_j}^{-1}$, where each $s_j \in \{0, 1\}$;
3. Let $\mathbf{S} = \text{diag}\{s_1, \ldots, s_n\}$. Set $\mathbf{U} = \tilde{\mathbf{U}}\mathbf{S}, \mathbf{D} = \tilde{\mathbf{D}}\mathbf{S}, \mathbf{V} = \tilde{\mathbf{V}}\mathbf{S}$, $K = \sum s_j$.

Parameterizing a set of nested models with binary variables has been a useful technique in a variety of contexts, including variable selection in regression models (Mitchell and Beauchamp 1988). We continue with this formulation because it allows for the construction of a relatively straightforward Gibbs sampling scheme to generate samples from the posterior distribution.

The matrices $\mathbf{U}, \mathbf{D}$ and $\mathbf{V}$ described in 1, 2 and 3 above are exchangeable under simultaneous permutation of their columns. It follows from Proposition 1 that, conditional on $s_1, \ldots, s_n$, the non-zero columns of $\mathbf{U}$ and $\mathbf{V}$ are random samples from the uniform distributions on $\mathcal{V}_{\sum s_j, m}$ and $\mathcal{V}_{\sum s_j, n}$ respectively, and that conditional on $\{s_j = 1, \mathbf{U}_{[,-j]}, \mathbf{V}_{[,-j]}\}$,

$$\mathbf{U}_{[,j]} \stackrel{d}{=} \mathbf{N}^u_{\{-j\}}\mathbf{u}, \qquad \mathbf{V}_{[,j]} \stackrel{d}{=} \mathbf{N}^v_{\{-j\}}\mathbf{v}, \text{ where}$$

- $\mathbf{N}^u_{\{-j\}}$ and $\mathbf{N}^v_{\{-j\}}$ are orthonormal bases for the null spaces of $\mathbf{U}_{[,-j]}$ and $\mathbf{V}_{[,-j]}$;
- $\mathbf{u}$ and $\mathbf{v}$ are uniformly distributed on the $(m - \sum s_j + 1)$- and $(n - \sum s_j + 1)$-spheres.

This property will facilitate posterior sampling of the columns of $\mathbf{U}$, $\mathbf{D}$ and $\mathbf{V}$, as described in the next subsection.

4.2 Posterior estimation

Let $\mathbf{\Theta} = \{\mathbf{U}, \mathbf{D}, \mathbf{V}\}$, $\mathbf{\Theta}_j = \{\mathbf{U}_{[,j]}, d_j, \mathbf{V}_{[,j]}\}$ and $\mathbf{\Theta}_{-j} = \{\mathbf{\Theta}_k : k \neq j\}$. In this subsection we derive the full conditional distribution of $\mathbf{\Theta}_j$ given $\{\mathbf{Y}, \mathbf{\Theta}_{-j}, \phi, \mu, \psi\}$ under the model described in the previous subsection. The prior and full conditional distributions of ϕ, μ and ψ remain unchanged from Section 2. The full conditional distributions can be used in a Gibbs sampling scheme to generate approximate samples from $p(\mathbf{U}, \mathbf{D}, \mathbf{V}, \phi, \lambda | \mathbf{Y})$.

Under the model and parameterization described above, the components of $\mathbf{\Theta}_j$ are either all zero or have a distribution as described in Section 2. To sample $\mathbf{\Theta}_j$, we first sample whether or

not the components are zero, and if not, sample the non-zero values. More specifically, sampling $\mathbf{\Theta}_j$ from its full conditional distribution can be achieved as follows:

1. Sample from $(\{d_j = 0\}, \{d_j \neq 0\})$ conditional on $\mathbf{Y}, \mathbf{\Theta}_{-j}, \phi, \mu, \psi$.
2. If $\{d_j = 0\}$ is true, then set $d_j, \mathbf{U}_{[,j]}$ and $\mathbf{V}_{[,j]}$ all equal to zero.
3. If $\{d_j \neq 0\}$ is true,
   (a) sample $d_j|\mathbf{Y}, \mathbf{\Theta}_{-j}, \phi, \mu, \psi, \{d_j \neq 0\}$;
   (b) sample $\{\mathbf{U}_{[,j]}, \mathbf{V}_{[,j]}\}|\mathbf{Y}, \mathbf{\Theta}_{-j}, \phi, d_j$.

The steps 1, 2, and 3 above constitute a draw from $p(\mathbf{\Theta}_j|\mathbf{Y}, \mathbf{\Theta}_{-j}, \phi, \mu, \psi)$. The first step requires calculation of the odds:

$$\text{odds}(d_j \neq 0|\mathbf{Y}, \mathbf{\Theta}_{-j}, \phi, \mu, \psi) = \frac{p(d_j \neq 0|\mathbf{\Theta}_{-j})}{p(d_j = 0|\mathbf{\Theta}_{-j})} \times \frac{p(\mathbf{Y}|\mathbf{\Theta}_{-j}, d_j \neq 0, \phi, \mu, \psi)}{p(\mathbf{Y}|\mathbf{\Theta}_{-j}, d_j = 0, \phi, \mu, \psi)} \tag{2}$$

The first ratio is simply the prior conditional odds of $\{d_j \neq 0\}$ and can be derived from the prior distribution on the rank K. The second term in (2) can be viewed as a Bayes factor, evaluating the evidence in the data for additional structure in $E[\mathbf{Y}]$ beyond that provided by $\mathbf{\Theta}_{-j}$. Recall from the previous section that $\mathbf{Y} - \mathbf{UDV}' = \mathbf{E}_{-j} - d_j\mathbf{U}_{[,j]}\mathbf{V}'_{[,j]}$, and so we can write

$$\begin{aligned} p(\mathbf{Y}|\mathbf{U}, \mathbf{D}, \mathbf{V}, \phi, \mu, \psi) &= \left[\left(\frac{\phi}{2\pi}\right)^{mn/2} \exp\{-\frac{1}{2}\phi||\mathbf{E}_{-j}||^2\}\right] \exp\{-\frac{1}{2}\phi d_j^2\} \exp\{\phi d_j \mathbf{U}'_{[,j]}\mathbf{E}_{-j}\mathbf{V}_{[,j]}\} \\ &= p(\mathbf{Y}|\mathbf{\Theta}_{-j}, d_j = 0, \phi, \mu, \psi) \times \exp\{-\frac{1}{2}\phi d_j^2\} \times \exp\{\phi d_j \mathbf{U}'_{[,j]}\mathbf{E}_{-j}\mathbf{V}_{[,j]}\} \end{aligned} \tag{3}$$

The first term in (3) is equal to the denominator of the Bayes factor, and is simply a product of normal densities with the elements of $\mathbf{Y}$ having means given by $\mathbf{U}_{[,-j]}\mathbf{D}_{[-j,-j]}\mathbf{V}'_{[,-j]}$ and equal variances $1/\phi$. The numerator of the Bayes factor can be obtained by integrating (3) over $\mathbf{\Theta}_j$ with respect to its conditional distribution given $\mu, \psi, \mathbf{\Theta}_{-j}$ and $\{d_j \neq 0\}$. Integrating first with respect to $\mathbf{U}_{[,j]}, \mathbf{V}_{[,j]}$, we need to calculate $E[\exp\{\phi d_j \mathbf{U}'_{[,j]}\mathbf{E}_{-j}\mathbf{V}_{[,j]}\}|\mathbf{\Theta}_{-j}, d_j]$. Let $\tilde{m} = m - \sum_{k\neq j}\{d_k \neq 0\}$ and $\tilde{n} = n - \sum_{k\neq j}\{d_k \neq 0\}$. Recall that conditional on $\mathbf{\Theta}_{-j}$, $\mathbf{U}_{[,j]} \stackrel{d}{=} \mathbf{N}^u_{\{-j\}}\mathbf{u}$ and $\mathbf{V}_{[,j]} \stackrel{d}{=} \mathbf{N}^v_{\{-j\}}\mathbf{v}$ where $\mathbf{u}$ and $\mathbf{v}$ are uniformly distributed on the $\tilde{m}$- and $\tilde{n}$-spheres. Letting $\tilde{\mathbf{E}} = \mathbf{N}^{u\prime}_{\{-j\}}\mathbf{E}_{-j}\mathbf{N}^v_{\{-j\}}$, the required expectation can therefore be rewritten as $\mathrm{E}_{\mathbf{uv}}[\exp\{\phi d_j \mathbf{u}'\tilde{\mathbf{E}}\mathbf{v}\}]$. This expectation is non-standard, and is derived in the appendix. The result gives:

$$p(\mathbf{Y}|\mathbf{\Theta}_{-j}, \phi, \mu, \psi, d_j) = p(\mathbf{Y}|\mathbf{\Theta}_{-j}, \phi, \mu, \psi, d_j = 0) \times \exp\{-\frac{1}{2}\phi d_j^2\} \sum_{l=0}^{\infty} ||\tilde{\mathbf{E}}||^{2l}\phi^{2l}d_j^{2l}a_l \tag{4}$$

where the sequence $\{a_l\}_0^\infty$ can be computed exactly and is given in the appendix.

The calculation of $p(\mathbf{Y}|\mathbf{\Theta}_{-j}, \phi, \mu, \psi, d_j \neq 0)$ is completed by integrating (4) over d_j with respect to $p(d_j|\mathbf{\Theta}_{-j}, \phi, \mu, \psi, d_j \neq 0)$, the normal density with mean μ and precision ψ. This integration simply requires calculating the even moments of a normal distribution, resulting in

$$p(\mathbf{Y}|\mathbf{\Theta}_{-j}, \phi, \psi, d_j \neq 0) = p(\mathbf{Y}|\mathbf{\Theta}_{-j}, \phi, \psi, d_j = 0) \times \sum_{l=0}^{\infty} ||\tilde{\mathbf{E}}||^{2l} a_l b_l \tag{5}$$

where the sequence $\{b_l\}_0^\infty$ is given by

$$b_l = \phi^{2l} \left(\frac{\psi}{\phi+\psi}\right)^{1/2} \exp\{-\frac{1}{2}\mu^2\psi\phi/(\phi+\psi)\} E[\{\frac{1}{\sqrt{\phi+\psi}}(Z + \frac{\mu\psi}{\phi+\psi})\}^{2l}]$$

where Z is standard normal. The required moments can be calculated iteratively, see for example Smith (1995). The conditional odds of $\{d_j \neq 0\}$ is therefore

$$\text{odds}(d_j \neq 0|\mathbf{Y}, \mathbf{\Theta}_{-j}, \phi, \lambda) = \frac{p(d_j \neq 0|\mathbf{\Theta}_{-j})}{p(d_j = 0|\mathbf{\Theta}_{-j})} \times \sum_{l=0}^{\infty} ||\tilde{\mathbf{E}}||^{2l} a_l b_l.$$

In practice, only a finite number of terms can be used to compute the above sums. The sum in (4) can be bounded above and below by modified Bessel functions, and the error in a finite-sum approximation can be bounded, at least to the extent that one can compute the bounding Bessel functions. This can also provide a guide as to how many terms to include in approximating (5). Details are given in the Appendix.

If $\{d_j \neq 0\}$ is sampled it is still necessary to sample $d_j, \mathbf{U}_{[,j]}$ and $\mathbf{V}_{[,j]}$. Multiplying equation (4) by the prior for $d_j|\{d_j \neq 0\}$, the required conditional distribution for d_j is proportional to

$$p(d_j|\mathbf{Y}, \mathbf{\Theta}_{-j}, \phi, \mu, \psi, \{d_j \neq 0\}) \propto e^{-\frac{1}{2}(d_j-\mu)^2\psi} e^{-\frac{1}{2}d_j^2\phi} \sum_{l=0}^{\infty} ||\tilde{\mathbf{E}}||^{2l} \phi^{2l} d_j^{2l} a_l$$

which is an infinite mixture with the following components:

- mixture weights: $w_l \propto ||\tilde{\mathbf{E}}||^{2l} a_l b_l$
- mixture densities: $f_l(d) \propto d^{2l} \exp\{-\frac{1}{2}(d-\tilde{\mu})^2\tilde{\psi}\}$, where $\tilde{\mu} = \mu\psi/(\phi+\psi)$ and $\tilde{\psi} = \phi + \psi$

The density $f_l(d)$ is nonstandard, but can be sampled from quite efficiently using rejection sampling with a scaled and shifted t-distribution as the approximating density (the tails of a normal distribution are not heavy enough).

To sample $\mathbf{U}_{[,j]}$ and $\mathbf{V}_{[,j]}$ we first sample $\mathbf{u}$ and $\mathbf{v}$ from their joint distribution and then set $\mathbf{U}_{[,j]} = \mathbf{N}^u_{\{-j\}}\mathbf{u}$ and $\mathbf{V}_{[,j]} = \mathbf{N}^v_{\{-j\}}\mathbf{v}$. Equation (3) indicates that the joint conditional density of $\{\mathbf{u}, \mathbf{v}\}$ is of the form

$$p(\mathbf{u}, \mathbf{v}|\mathbf{Y}, \mathbf{\Theta}_{-j}, \phi, \mu, \psi, d_j) = c(\mathbf{A}) \exp\{\mathbf{u}'\mathbf{A}\mathbf{v}\}, \tag{6}$$

where $\mathbf{A} = \phi d_j \tilde{\mathbf{E}}$ and $c(\mathbf{A})^{-1} = c_{\tilde{m}}(0)^{-1} c_{\tilde{n}}(0)^{-1} \sum_{l=0}^{\infty} ||\mathbf{A}||^{2l} a_l$. This density defines a joint distribution for two dependent unit vectors. To my knowledge, such a joint distribution has not been studied before. Some useful facts about this distribution are

- the conditional distribution of $\mathbf{u}|\mathbf{v}$ is vMF($\mathbf{Av}$), and that of $\mathbf{v}|\mathbf{u}$ is vMF($\mathbf{u}'\mathbf{A}$);
- the marginal distribution of $\mathbf{v}$ is proportional to $I_{\tilde{m}/2-1}(||\mathbf{Av}||)/||\mathbf{Av}||^{\tilde{m}/2-1}$;
- the joint density has local maxima at $\{\pm(\hat{\mathbf{u}}_k, \hat{\mathbf{v}}_k), k = 1, \ldots, \tilde{n}\}$ where $(\hat{\mathbf{u}}_k, \hat{\mathbf{v}}_k)$ are the kth singular vectors of $\mathbf{A}$.

I have implemented a number of rejection and importance samplers for this distribution, although making these schemes efficient is still a work in progress. A relatively fast approximate method that seems to work well for a variety of matrices $\mathbf{A}$ is to first sample $(\mathbf{u}, \mathbf{v})$ from the local modes $\{\pm(\hat{\mathbf{u}}_k, \hat{\mathbf{v}}_k), k = 1, \ldots, \tilde{n}\}$ according to the exact relative probabilities and then use this value as a starting point for a small number of Gibbs samples, alternately sampling from $p(\mathbf{u}|\mathbf{A}, \mathbf{v})$ and $p(\mathbf{v}|\mathbf{A}, \mathbf{u})$.

The complexity of the calculations involved in sampling from $p(\mathbf{\Theta}_j|\mathbf{Y}, \mathbf{\Theta}_{-j}, \phi, \mu, \psi)$ suggest we look for a simpler procedure. For example, we could model only d_j to be zero with non-zero probability, and sample from its full conditional distribution instead of marginally over $\mathbf{U}_{[,j]}$ and $\mathbf{V}_{[,j]}$. Unfortunately, an algorithm based on this approach will not mix well across ranks of $\mathbf{M}$ because d_j, $\mathbf{U}_{[,j]}$ and $\mathbf{V}_{[,j]}$ are dependent to an extreme: The probability of sampling $d_j \neq 0$ is essentially zero unless $\mathbf{U}_{[,j]}$ and $\mathbf{V}_{[,j]}$ are near a pair of local modes, but the probability of $\mathbf{U}_{[,j]}$ and $\mathbf{V}_{[,j]}$ being in such a state is essentially zero if $d_j = 0$. Metropolis-Hastings algorithms are problematic for similar reasons and, based on my initial efforts on this problem, such algorithms seem to require an extreme amount of tuning of the proposal distributions to achieve even minimal acceptance rates. In contrast, sampling d_j marginal over $\mathbf{U}_{[,j]}$ and $\mathbf{V}_{[,j]}$ is possible as shown above, requires no tuning and, for the examples in this article, mixes well across matrices $\mathbf{M}$ of different ranks.

4.3 A suggested Gibbs sampling scheme

The dimension-changing Monte Carlo sampler outlined above is computationally expensive compared to the fixed-dimension sampler of Section 3. For this reason, it may be desirable to incorporate the fixed-dimension sampler even when we are interested in sampling across dimensions, as this might improve within-dimension mixing at a low computational cost. One such algorithm proceeds by iterating the following steps:

A. Variable dimension sampler: For each $j \in \{1, \ldots, n\}$, sample $\mathbf{\Theta}_j = \{\mathbf{U}_{[,j]}, d_j, \mathbf{V}_{[,j]}\}$ via

- sampling $d_j|\mathbf{Y}, \mathbf{\Theta}_{-j}, \phi, \mu, \psi$;
- sampling $(\mathbf{U}_{[,j]}, \mathbf{V}_{[,j]})|\mathbf{Y}, \mathbf{\Theta}_{-j}, \phi, \mu, \psi, d_j$.

B. Fixed dimension sampler: For each $\{j : d_j \neq 0\}$,

- sample $\mathbf{U}_{[,j]}|\mathbf{Y}, \mathbf{\Theta}_{-j}, \phi, \mu, \psi, d_j, \mathbf{V}_{[,j]}$;
- sample $\mathbf{V}_{[,j]}|\mathbf{Y}, \mathbf{\Theta}_{-j}, \phi, \mu, \psi, d_j, \mathbf{U}_{[,j]}$;
- sample $d_j|\mathbf{Y}, \mathbf{\Theta}_{-j}, \phi, \mu, \psi, \mathbf{U}_{[,j]}, \mathbf{V}_{[,j]}$.

C. Other terms:

- sample $\phi|\mathbf{Y}, \mathbf{\Theta}$;
- sample $\mu|\mathbf{D}, \psi$;
- sample $\psi|\mathbf{D}, \mu$.

Alternatively, steps **A** and **B** could be performed on random subsets of indices j. The distributions required for the steps in **A** are outlined in this section, and steps **B** and **C** are outlined in the previous section. By conditioning on whether or not $d_j = 0$ for each j, the steps in **B** can be viewed as Gibbs sampling for all $j \in \{1, \ldots, n\}$, not just those for which $d_j \neq 0$. R-code that implements this routine is available at my website.

5 Simulation study

In this section we examine the sampling properties of the estimation procedure with a small simulation study. Each dataset in this study was simulated from the following model:

- $\mathbf{U} \sim$ uniform$(\mathcal{V}_{5,m})$, $\mathbf{V} \sim$ uniform$(\mathcal{V}_{5,n})$;
- $\mathbf{D} = \text{diag}\{d_1, \ldots, d_5\}$, $\{d_1, \ldots, d_5\} \sim$ i.i.d. uniform$(\frac{1}{2}\mu_{mn}, \frac{3}{2}\mu_{mn})$.
- $\mathbf{Y} = \mathbf{UDV}' + \mathbf{E}$, where $\mathbf{E}$ is an $m \times n$ matrix of standard normal noise.

For each value of m and n, the sampling mean of $\{d_1, \ldots, d_5\}$ was taken to be $\mu_{mn} = \sqrt{n + m + 2\sqrt{nm}}$. Such a value should distribute the singular values $\{d_1, \ldots, d_5\}$ near the "cusp" of detectability: As shown in Edelman (1988), the largest singular value of an $m \times n$ matrix $\mathbf{E}$ of standard normal noise is approximately μ_{mn} for large m and n.

Three-hundred datasets were generated using the model above, one-hundred for each of the three sample sizes $(m, n) \in \{(10, 10), (100, 10), (100, 100)\}$. These were generated in the R statistical computing environment using the integers 1 through 100 as random seeds for each of the three sample sizes. Code to generate these datasets is available from my website. Prior distributions for the parameters $\{\phi, \mu, \psi\}$ were taken as described above with "prior sample sizes" of $\nu_0 = 2$ and $\eta_0 = 2$. This gives exponential prior distributions for ϕ and ψ. The values of σ_0^2, μ_0 and τ_0^2 were derived from "empirical Bayes"-type estimates obtained by averaging over different ranks as follows:

1. For each $k \in \{0, \ldots, n\}$,
   (a) Let $\hat{\mathbf{U}}\hat{\mathbf{D}}\hat{\mathbf{V}}'$ be the least-squares projection of $\mathbf{Y}$ onto the set of rank-k matrices;
   (b) Let $\hat{\sigma}_k^2 = ||\mathbf{Y} - \hat{\mathbf{U}}\hat{\mathbf{D}}\hat{\mathbf{V}}'||^2/(nm)$
   (c) Let $\hat{\mu}_k = \sum_{j=1}^k \hat{d}_j/k$, $\hat{\tau}_k^2 = \sum_{j-1}^k (\hat{d}_j - \bar{\hat{d}})^2/k$.
2. Let $\sigma_0^2 = \frac{1}{n+1}\sum_{j=0}^n \hat{\sigma}_j^2$, $\mu_0 = \frac{1}{n+1}\sum_{j=0}^n \hat{\mu}_j$, $v_0^2 = \frac{1}{n}\sum_{j=0}^n (\hat{\mu}_j - \bar{\hat{\mu}})^2$, $\tau_0^2 = \frac{1}{n+1}\sum_{j=0}^n \hat{\tau}_j^2$.

The resulting prior distributions are weakly centered around averages of empirical estimates, where the averaging is over ranks 0 through n. Finally, the prior distribution on the rank K of the mean matrix was taken to be uniform on $\{0, \ldots, n\}$. Other simple priors that gave similar results for this simulation study include diffuse mean-zero normal distributions for the d_j's (used in the next section), and one in which the conditional mean of the d_j's given ϕ is $\phi^{-1/2}\sqrt{m+n+2\sqrt{mn}}$. This latter prior distribution essentially focuses the search for non-zero d_j's to values that are as large as the largest singular values of normally distributed noise matrices, and will generally result in a posterior distribution that puts more mass on lower values of K than would a prior distribution for the d_j's centered around zero. A more complicated alternative to these approaches would be to have the prior distributions for $\{\phi, \mu, \psi\}$ depend on K. For example, given $K = k$, the prior distributions for $\{\phi, \mu, \psi\}$ could be based on $\{\hat{\sigma}_k^2, \hat{\mu}_k, \hat{\tau}_k^2\}$. Such prior distributions would require some minor modifications to the variable-dimension sampler outlined in the previous section.

For each of the 100×3 datasets, $10,000$ iterations of the Gibbs sampling scheme described in Section 4.3 were run to obtain approximate samples from the posterior distribution of $\mathbf{UDV}'$. All Markov chains were begun with $K = 0$ and $\{\phi, \mu, \psi\}$ set equal to their prior modes. Summaries of the posterior distributions for the three different values of (m, n) are displayed in Figure 2. The first column of each panel plots the MCMC approximation to the expected value of $p(K|\mathbf{Y})$ for each value of (m, n). The expectation $E_Y[p(K|\mathbf{Y})]$ is approximated by $\frac{1}{100}\sum_{s=1}^{100} p(K|\mathbf{Y}_{(s)})$, where $\mathbf{Y}_{(s)}$ is the sth simulated dataset for a given value of (m, n) (for the case $m = n = 100$, $p(K|\mathbf{Y})$ is plotted only for $K \leq 10$, although the distribution extends beyond this value). These distributions are all peaked around the correct value $K = 5$ Also of interest is how frequently the posterior mode $\hat{K} = \arg\max p(K|\mathbf{Y})$ obtains the true value of $K = 5$. This information is displayed in the second column of Figure 2, which gives the empirical distribution of $\hat{K}$ taken over each of the 100 datasets. As we see, the true value $K = 5$ is the most frequent value of the estimate in each dataset, with $K = 4$ a close second. This is not too surprising, as half of the simulated singular values are below the rough detection threshold of $\sqrt{n+m+2\sqrt{nm}}$.

Lastly we consider the effect of shrinkage on the estimate of the mean matrix $\mathbf{M}$. For each simulated dataset the posterior mean $\hat{\mathbf{M}} = E[\mathbf{M}|\mathbf{Y}]$ was obtained by averaging its value over the 10^4 scans

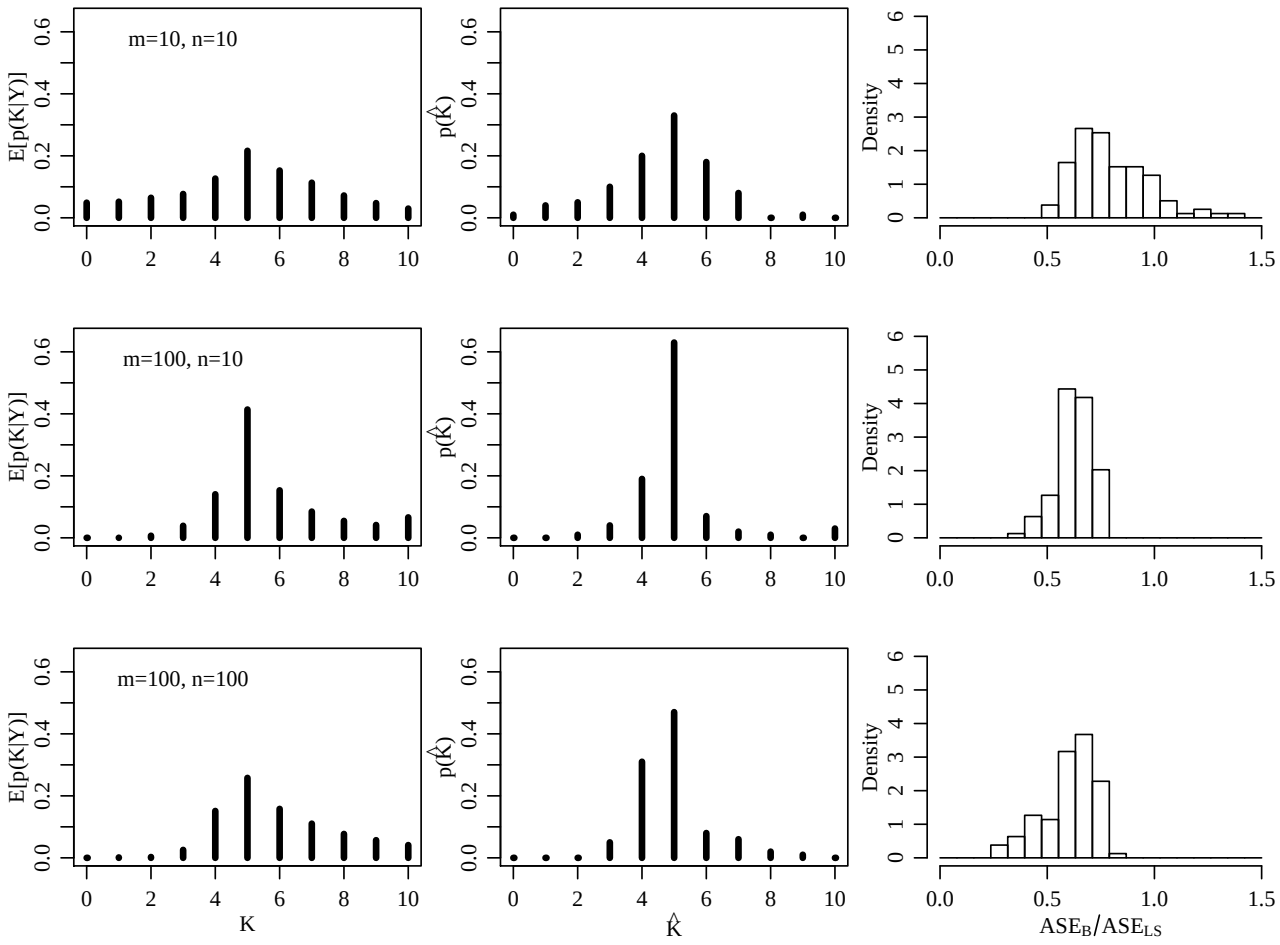

Figure 2: Results of the simulation study. Plots in the first column give the averages of $p(K|\mathbf{Y})$ over 100 simulated datasets. The second column gives the empirical distribution of the posterior mode $\hat{K}$. The third column gives the distribution of the ratio of the squared error of the Bayes estimate of $\mathbf{M}$ to that of the least-squares estimate.

of the Gibbs sampler. The squared error in estimation, averaged over elements of the mean matrix was calculated as $\text{ASE}_B = ||\hat{\mathbf{M}} - \mathbf{M}||^2/(mn)$ where $\mathbf{M}$ is the mean matrix that generated the data. This value is compared to ASE_{LS}, which is the corresponding average squared error of the least-squares projection of $\mathbf{Y}$ onto the space of rank-$\hat{K}$ matrices. The distribution of this ratio is mostly below 1 for the case $m = n = 10$, and strictly below 1 for the other two cases where there are more parameters to estimate. This corresponds with our intuition: The model-averaged estimates improve relative to the least-squares estimates as the number of parameters increases. These results indicate that simply obtaining a posterior estimate $\hat{K}$ of K and then using the corresponding rank-$\hat{K}$ least-squares estimate of $\mathbf{M}$ generally results in an estimate that can be substantially improved upon by model averaging, at least in terms of this error criterion.

6 Extension and example: analysis of binary relational data

A potentially useful extension of the model described above is to a class of generalized bilinear models of the form

$$\begin{aligned} \theta_{i,j} &= \boldsymbol{\beta}'\mathbf{x}_{i,j} + \mathbf{u}_i'\mathbf{D}\mathbf{v}_j + e_{i,j} \\ E[y_{i,j}|\boldsymbol{\Theta}] &= g^{-1}(\theta_{i,j}) \end{aligned}$$

where g is the link function. Such models allow for the analysis of a variety of data types: For example, binary data can be modeled as $y_{i,j} \sim \text{binary}(\frac{\exp\{\theta_{i,j}\}}{1+\exp\{\theta_{i,j}\}})$ and count data as $y_{i,j} \sim \text{Poisson}(\exp\{\theta_{i,j}\})$. Gabriel (1998) considered maximum likelihood estimation for a variant of this model in situations where the dimension of $\mathbf{D}$ is fixed, and Hoff (2005) considered a symmetric version of this model for the analysis of social network data. Parameter estimation and dimension selection for the above model can be made by sampling from a Markov chain generated by a modified version of the algorithm of Section 4.3. Given current values of $\boldsymbol{\Theta}, \boldsymbol{\beta}, \mathbf{U}, \mathbf{D}, \mathbf{V}$, sample new values as follows:

1. Let $\tilde{\mathbf{Y}} = \boldsymbol{\Theta} - \mathbf{X}\boldsymbol{\beta} = \mathbf{U}\mathbf{D}\mathbf{V}' + \mathbf{E}$. Update $\mathbf{U}$, $\mathbf{D}$, and $\mathbf{V}$ from their conditional distribution given $\tilde{\mathbf{Y}}$ as described in Section 4.3.

2. Let $\tilde{\mathbf{Y}} = \boldsymbol{\Theta} - \mathbf{U}\mathbf{D}\mathbf{V}' = \mathbf{X}\boldsymbol{\beta} + \mathbf{E}$. Update $\boldsymbol{\beta}$ from its conditional distribution given $\tilde{\mathbf{Y}}$ (a multivariate normal distribution).

3. Sample $\boldsymbol{\Theta}^* = \mathbf{X}\boldsymbol{\beta} + \mathbf{U}\mathbf{D}\mathbf{V}' + \mathbf{E}^*$, where $\mathbf{E}^*$ is a matrix of normally distributed noise with zero mean and precision ϕ. Replace $\theta_{i,j}$ by $\theta^*_{i,j}$ with probability $\frac{p(y_{i,j}|\theta^*_{i,j})}{p(y_{i,j}|\theta_{i,j})} \wedge 1$.

We illustrate the use of such a model and estimation procedure with an analysis of binary relational data between 46 global service firms and 55 cities, obtained from the Globalization and

World Cities study group (`http://www.lboro.ac.uk/gawc`). For these data, $y_{i,j} = 1$ if firm j has an office in city i and $y_{i,j} = 0$ otherwise. Standard practice is to represent within-row and within-column homogeneity with effects that are additive on the log-odds scale:

$$\log \text{odds}(y_{i,j} = 1) = \beta + a_i + b_j, \tag{7}$$

and so the effects $\mathbf{a} = \{a_1, \ldots, a_m\}$ and $\mathbf{b} = \{b_1, \ldots, b_n\}$ constitute a rank-two structure. We look for evidence of higher-order structure by considering the model

$$\begin{aligned} \log \text{odds}(y_{i,j} = 1) &= \beta + \gamma_{i,j} \\ \gamma_{i,j} &= \mathbf{u}_i{}'\mathbf{D}\mathbf{v}_j + e_{i,j} \end{aligned} \tag{8}$$

The rank-two structure of model (7) is easily incorporated into (8) by fixing $\mathbf{U}_{[,1]} = \frac{1}{\sqrt{m}}\mathbf{1}_{m\times 1}$ and $\mathbf{V}_{[,2]} = \frac{1}{\sqrt{n}}\mathbf{1}_{n\times 1}$ and modeling d_1 and d_2 to be non-zero with probability 1. The additive city and firm effects are then given by $\mathbf{a} = d_2\mathbf{U}_{[,2]}$ and $\mathbf{b} = d_1\mathbf{V}_{[,1]}$ respectively. Note that any remaining effects represented by $\mathbf{UDV}'$ will be orthogonal to these additive effects, and that the mean of the matrix $\mathbf{UDV}'$ is identically zero, making it unaliased with the intercept β. For the remainder of this analysis, the variable K will refer to the number of additional non-zero singular values of $\mathbf{UDV}'$ beyond the additive row and column effects.

We fix the error variance $1/\phi = 1$, as this scaling parameter is confounded with the magnitude of β and $\mathbf{UDV}'$. For simplicity we use independent normal $(0, 100)$ prior distributions for β and the non-zero elements of $\mathbf{D}$, and a uniform prior distribution for K. A Markov chain of length 25,000 was constructed using the algorithm described above, starting with $K = 0$. Mixing across ranks K was quite rapid as is shown in the first panel of Figure 3, which displays values of K every 100th scan of the Markov chain. The Monte Carlo estimate of $p(K|\mathbf{Y})$, shown in the second panel, gives a posterior mode of $K = 6$ and suggests strong evidence for structure in the log-odds beyond that of the additive row and column effects.

One of the practical motivations for selecting an appropriate model dimension is prediction. Many binary social network datasets include missing values, in which it is not known whether $y_{i,j} = 1$ or $y_{i,j} = 0$. In such cases it is often desirable to make predictions about missing values based on the observed data, and thus to base model selection on predictive performance. With this in mind, we compare the above results to the following 10-fold cross validation procedure:

1. Randomly split the set of pairs $\{i, j\}$ into ten test sets $A_1, \ldots, A_{10}$.
2. For $K = 0, 1, \ldots, K_{\max}$:
   (a) For $l = 1, \ldots, 10$:
      i. With the rank fixed at K, perform the MCMC algorithm using only $\{y_{i,j} : \{i, j\} \notin A_l\}$, but sample values of $\theta_{i,j}$ for all ordered pairs.

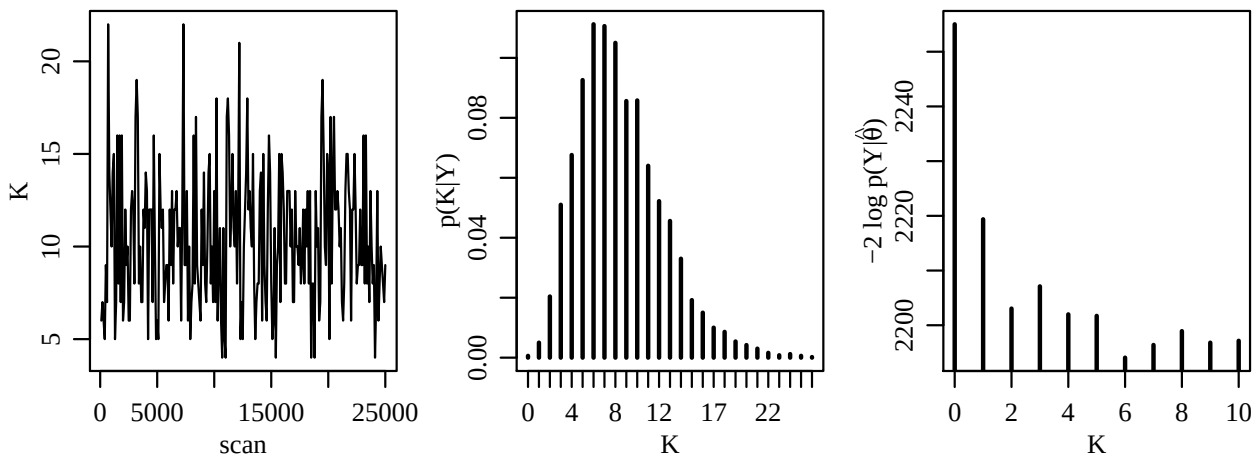

Figure 3: Posterior estimation of K. The first panel plots values of K every 100th scan of the Markov chain. The second panel plots the Monte Carlo estimate of $p(K|\mathbf{Y})$. The third panel gives the results of a cross-validation evaluation of $K \in \{0, \ldots, 10\}$.

ii. Based on the Monte Carlo sample values $\{\theta_{i,j}^{(1)}, \ldots, \theta_{i,j}^{(S)}\}$ compute the posterior mean $\hat{\mu}_{i,j} = \frac{1}{S}\sum_{s=1}^{S} \frac{\exp\{\theta_{i,j}^{(s)}\}}{1+\exp\{\theta_{i,j}^{(s)}\}}$ for $\{i,j\} \in A_l$ and the log predictive probability $\text{lpp}(A_l) = \sum_{\{i,j\}\in A_l} \log p(y_{i,j}|\hat{\mu}_{i,j})$.

(b) Measure the predictive performance for K as $\text{LPP}(K) = \sum_{l=1}^{K_{\max}} \text{lpp}(A_l)$.

The values of $-2\text{LPP}(K)$ for $K \in \{0, \ldots, 10\}$ are shown in the third panel of Figure (3). For the particular random partitioning of the data used here, the cross-validation procedure suggests a model rank of $K = 6$, which is the same value as the posterior mode of the Bayes solution. However, a comparison of N values of K using a ten-fold cross validation procedure requires the construction of $10 \times N$ separate Markov chains, and further requires specification of the values of K to be compared. In contrast, the Bayesian procedure requires only one MCMC run and can potentially visit each value of $K \in \{1, \ldots, n\}$.

Finally we examine some of the patterns in the structure of $\mathbf{UDV}'$ beyond those of the additive effects. The posterior mean of $\mathbf{UDV}$, minus the additive effects, was obtained by averaging over scans of the Markov chain. The first two singular values and vectors of this matrix were obtained, and the values of the resulting row (city) effects are plotted in Figure 4. These values are strongly related to geography: U.S. cities cluster together, as do cities in Europe, Latin America and from the Pacific rim.

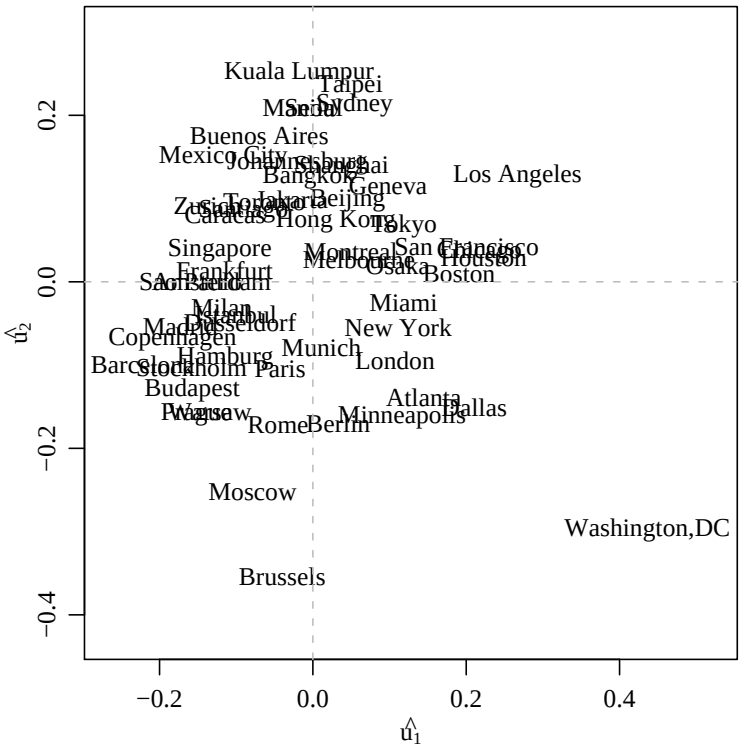

Figure 4: City specific effects: The first two left singular vectors of $E(M|\mathbf{Y})$ indicate strong geographic patterns in the data.

7 Discussion

This paper has presented a model-based data reduction and representation method for multivariate and matrix-valued data. The approach is to model the data matrix $\mathbf{Y}$ as equal to a reduced-rank mean matrix $\mathbf{M}$ plus Gaussian noise, and to simultaneously estimate $\mathbf{M}$ along with its rank. The approach is Bayesian and the estimation procedure, based on Markov chain Monte Carlo, allows for a wide variety of model extensions, such as to generalized bilinear models as described in the previous section. Other straightforward extensions include estimation using replicate data matrices and estimation subject to missing data. This latter extension may be of particular use in the analysis of relational data among a large number of nodes, where it may be too costly to make observations on all possible pairs. In such cases, the value of $y_{i,j}$ may be missing for many pairs, but one can make predictions based on estimates $\mathbf{u}_i, \mathbf{D}, \mathbf{v}_j$ obtained from the observed data. Using this approach to predict missing links in social networks and protein-protein interaction networks is one of my current research areas. However, for large datasets with 1000 nodes (10^6 observations) or more, the MCMC scheme in this article becomes prohibitively computationally expensive. I am currently studying methods of making approximate Bayesian inference for large relational datasets. These include Laplace approximations for various components of the MCMC scheme of Sections 3 and 4, and using variational methods for approximating joint posterior distributions (Jordan et al., 1999).

Computer code and data for all numerical results in this paper are available at `www.stat.washington.edu/hoff`.

A Proof of proposition 1

We first construct a sample from the uniform distribution on $\mathcal{V}_{K,m}$ and then show that it has the desired conditional distributions. Let $\mathbf{z}_1, \ldots, \mathbf{z}_K$ be i.i.d. multivariate normal $(0, \mathbf{I}_{m\times m})$. Let $\mathbf{x}_1 = \mathbf{z}_1$ and for $j = 1, \ldots, K-1$ let

- $\mathbf{X}_j = (\mathbf{x}_1 \cdots \mathbf{x}_j)$;
- $\mathbf{P}_j = \mathbf{I} - \mathbf{X}_j(\mathbf{X}_j'\mathbf{X}_j)^{-1}\mathbf{X}_j'$;
- $\mathbf{x}_{j+1} = \mathbf{P}_j\mathbf{z}_{j+1}$.

Note that $\mathbf{P}_j$ is the symmetric, idempotent projection matrix of $\mathbb{R}^K$ onto the null space of $\mathbf{X}_j$, and so the vectors $\mathbf{x}_1, \ldots, \mathbf{x}_{j+1}$ are orthogonal. For each j, let $\mathbf{U}_j = \mathbf{X}_j(\mathbf{X}_j'\mathbf{X}_j)^{-1/2}$. For $j = K$, we

have

$$\mathbf{X}_K'\mathbf{X}_K = \begin{pmatrix} \mathbf{x}_1' \\ \mathbf{x}_2' \\ \vdots \\ \mathbf{x}_K \end{pmatrix} (\mathbf{x}_1\ \mathbf{x}_2\ \cdots\ \mathbf{x}_K) = \begin{pmatrix} |\mathbf{x}_1|^2 & 0 & \cdots & 0 \\ 0 & |\mathbf{x}_2|^2 & \cdots & 0 \\ 0 & 0 & \cdots & |\mathbf{x}_K|^2 \end{pmatrix}$$

and so

$$\mathbf{U}_K = (\frac{\mathbf{x}_1}{|\mathbf{x}_1|},\ \frac{\mathbf{x}_2}{|\mathbf{x}_2|}, \cdots,\ \frac{\mathbf{x}_K}{|\mathbf{x}_K|})$$

is a matrix of K orthonormal vectors in $\mathbb{R}^m$. The proof will be complete if we can show the following:

Lemma 1: The distribution of $\mathbf{U}_K$ is the uniform distribution on $\mathcal{V}_{K,m}$.

Lemma 2: $\mathbf{U}_{[,k+1]}|\mathbf{U}_k \stackrel{d}{=} \mathbf{N}_k\mathbf{u}_{k+1}$ where $\mathbf{N}_k$ is an orthonormal basis for the null space of $\mathbf{U}_k$ and $\mathbf{u}_{k+1}$ is distributed uniformly on the $m-k$ dimensional sphere.

Proof of Lemma 1. By Theorem 2.2.1 (iii) of Chikuse (2003), an $m \times K$ matrix of the form $\mathbf{X}_K(\mathbf{X}_K'\mathbf{X}_K)^{-1/2}$ is uniformly distributed on $\mathcal{V}_{K,m}$ if $\mathbf{X}_K$ is an $m \times K$ random matrix with rank K a.s. and having a distribution that is invariant under left-orthogonal transformations. We will show left invariance for each $\mathbf{X}_k$ constructed above by induction. Let $\mathbf{H} : \mathbb{R}^m \to \mathbb{R}^m$ be an orthogonal transformation, and note that $\mathbf{HX}_1 = \mathbf{Hx}_1 \stackrel{d}{=} \mathbf{x}_1 = \mathbf{X}_1$. Now suppose $\mathbf{HX}_k \stackrel{d}{=} \mathbf{X}_k$. The distribution of $\mathbf{HX}_{k+1}$ is determined by its characteristic function:

$$E[\exp\{i\sum_{j=1}^{k+1}\mathbf{t}_j'\mathbf{H}\mathbf{x}_j\}] \quad = \quad E[\exp\{i\sum_{j=1}^{k}\mathbf{t}_j'\mathbf{H}\mathbf{x}_j\}E[\exp\{i\mathbf{t}_{k+1}'\mathbf{H}\mathbf{x}_{k+1}\}|\mathbf{X}_k]]$$

Note that $\mathbf{t}_{k+1}'\mathbf{H}\mathbf{x}_{k+1} = (\mathbf{P}_k'\mathbf{H}'\mathbf{t}_{k+1})'\mathbf{z}_{k+1}$, where $\mathbf{z}_{k+1}$ is a vector of independent standard normals and independent of $\mathbf{X}_k$. Thus the characteristic function can be rewritten as

$$E[\exp\{i\sum_{j=1}^{k}\mathbf{t}_j'\mathbf{H}\mathbf{x}_j\}\exp\{-\frac{1}{2}\mathbf{t}_{k+1}'\mathbf{H}\mathbf{P}_k\mathbf{P}_k'\mathbf{H}'\mathbf{t}_{k+1}\}] = E[\exp\{i\sum_{j=1}^{k}\mathbf{t}_j'\tilde{\mathbf{x}}_j\}\exp\{-\frac{1}{2}\mathbf{t}_{k+1}'\tilde{\mathbf{P}}_k\mathbf{t}_{k+1}\}] \quad (9)$$

where $\tilde{\mathbf{x}}_j = \mathbf{H}\mathbf{x}_j$ and

$$\begin{aligned} \tilde{\mathbf{P}}_k = \mathbf{H}\mathbf{P}_k\mathbf{P}_k'\mathbf{H}' &= \mathbf{H}\mathbf{P}_k\mathbf{H}' \\ &= \mathbf{H}(\mathbf{I} - \mathbf{X}_k(\mathbf{X}_k'\mathbf{X}_k)^{-1}\mathbf{X}_k')\mathbf{H}' \\ &= \mathbf{I} - \mathbf{H}\mathbf{X}_k((\mathbf{H}\mathbf{X}_k)'(\mathbf{H}\mathbf{X}_k))^{-1}(\mathbf{H}\mathbf{X}_k)' \end{aligned}$$

A similar calculation shows that the distribution of $\mathbf{X}_j$ is characterized by

$$E[\exp\{i\sum_{j=1}^{k+1}\mathbf{t}_j'\mathbf{x}_j\}] = E[\exp\{i\sum_{j=1}^{k}\mathbf{t}_j'\mathbf{x}_j\}\exp\{-\frac{1}{2}\mathbf{t}_{k+1}'\mathbf{P}_k\mathbf{t}_{k+1}\}], \quad (10)$$

By assumption, $\mathbf{X}_k \stackrel{d}{=} \mathbf{H}\mathbf{X}_k$, and so $\{\mathbf{x}_1, \ldots, \mathbf{x}_k, \mathbf{P}_k\} \stackrel{d}{=} \{\tilde{\mathbf{x}}_1, \ldots, \tilde{\mathbf{x}}_k, \tilde{\mathbf{P}}_k\}$ and the expectations (9) and (10) are equal. Since the characteristic functions specify the distributions, $\mathbf{H}\mathbf{X}_{k+1} \stackrel{d}{=} \mathbf{X}_{k+1}$ and the lemma is proved.

Proof of Lemma 2: The vector $\mathbf{U}_{[,k+1]}$ is constructed as $\mathbf{U}_{[,k+1]} = \mathbf{P}_k\mathbf{z}_{k+1}/|\mathbf{P}_k\mathbf{z}_{k+1}|$. $\mathbf{P}_k$ has $m-k$ eigenvalues of one, the rest being zero, giving the eigenvalue decomposition $\mathbf{P}_k = \mathbf{N}_k\mathbf{N}_k'$ where $\mathbf{N}_k$ is a $m \times (m-k)$ matrix whose columns form an orthonormal basis for the null space of $\mathbf{U}_k$. Substituting in $\mathbf{N}_k\mathbf{N}_k'$ for $\mathbf{P}_k$ gives

$$\begin{aligned}
\mathbf{U}_{[,k+1]} &= \frac{\mathbf{N}_k\mathbf{N}_k'\mathbf{z}_{k+1}}{|\mathbf{N}_k\mathbf{N}_k'\mathbf{z}_{k+1}|} \\
&= \mathbf{N}_k \frac{\mathbf{N}_k'\mathbf{z}_{k+1}}{(\mathbf{z}'\mathbf{N}_k\mathbf{N}_k'\mathbf{N}_k\mathbf{N}_k'\mathbf{z})^{1/2}} \\
&= \mathbf{N}_k \frac{\mathbf{N}_k'\mathbf{z}_{k+1}}{(\mathbf{z}'\mathbf{N}_k\mathbf{N}_k'\mathbf{z})^{1/2}} \\
&= \mathbf{N}_k \frac{\mathbf{N}_k'\mathbf{z}}{|\mathbf{N}_k'\mathbf{z}|}
\end{aligned}$$

Note that for each k, $\mathbf{U}_k = \mathbf{X}_k(\mathbf{X}_k'\mathbf{X}_k)^{-1/2}$, and so the projection matrix $\mathbf{P}_k$ can be written as $\mathbf{I} - \mathbf{U}_k\mathbf{U}_k'$, a function of $\mathbf{U}_k$. Therefore, given $\mathbf{U}_k$, $\mathbf{U}_{[,k+1]}$ is equal in distribution to $\mathbf{N}_k$ (a function of $\mathbf{U}_k$) multiplied by $\mathbf{N}_k'\mathbf{z}/|\mathbf{N}_k'\mathbf{z}|$. The distribution of $\mathbf{N}_k'\mathbf{z}$ can be found via its characteristic function: For an $m-k$-vector $\mathbf{t}$

$$\begin{aligned}
E[\exp\{i\mathbf{t}'(\mathbf{N}_k'\mathbf{z})\}] &= E[\exp\{i(\mathbf{N}_k\mathbf{t})'\mathbf{z}\}] \\
&= \exp\{-\frac{1}{2}\mathbf{t}'\mathbf{N}_k'\mathbf{N}_k\mathbf{t}\} \\
&= \exp\{-\frac{1}{2}\mathbf{t}'\mathbf{t}\},
\end{aligned}$$

and so we see that $\mathbf{N}_k\mathbf{z}_{k+1}$ is equal in distribution to an $m-k$-vector of independent standard normal random variables, and so $\mathbf{N}_k\mathbf{z}_{k+1}/|\mathbf{N}_k\mathbf{z}_{k+1}|$ is uniformly distributed on on the $m-k$-sphere.

B Expectation of $e^{\mathbf{u}'\mathbf{A}\mathbf{v}}$

In this section we compute $E[e^{\mathbf{u}'\mathbf{A}\mathbf{v}}]$ for uniformly distributed unit vectors $\mathbf{u}$ and $\mathbf{v}$ and an arbitrary $m \times n$ matrix $\mathbf{A}$. Integrating with respect to $\mathbf{v}$ can be accomplished by noting that as a function of $\mathbf{v}$, $e^{\mathbf{u}'\mathbf{A}\mathbf{v}}$ is proportional to the von Mises-Fisher distribution on the n-sphere S_n, with parameter $\mathbf{u}'\mathbf{A}$:

$$\begin{aligned}
\int e^{\mathbf{u}'\mathbf{A}\mathbf{v}} p(\mathbf{v})\ dS_n(\mathbf{v}) &= \int e^{\mathbf{u}'\mathbf{A}\mathbf{v}} c_n(0)\ dS_n(\mathbf{v}) \\
&= \frac{c_n(0)}{c_n(||\mathbf{u}'\mathbf{A}||)} \int e^{\mathbf{u}'\mathbf{A}\mathbf{v}} c_n(||\mathbf{u}'\mathbf{A}||)\ dS_n(\mathbf{v}) \\
&= \frac{c_n(0)}{c_n(||\mathbf{u}'\mathbf{A}||)} \\
&= \Gamma(n/2)(2/||\mathbf{u}'\mathbf{A}||)^{n/2-1} I_{n/2-1}(||\mathbf{u}'\mathbf{A}||)
\end{aligned}$$

where I_ν is the modified Bessel function of the first kind. The series expansion of $I_{n/2-1}(||\mathbf{u}'\mathbf{A}||)$ gives

$$\Gamma(n/2)(2/||\mathbf{u}'\mathbf{A}||)^{n/2-1} I_{n/2-1}(||\mathbf{u}'\mathbf{A}||) = \sum_{l=0}^{\infty} ||\mathbf{u}'\mathbf{A}||^{2l} \frac{\Gamma(n/2)}{\Gamma(l+1)\Gamma(l+n/2)4^l}.$$

All the terms in the sum are positive, so $E[e^{\mathbf{u}'\mathbf{A}\mathbf{v}}]$ can be found by replacing $||\mathbf{u}'\mathbf{A}||^{2l}$ with its expectation in the above equation. To compute this expectation, let $\mathbf{A} = \mathbf{L}\mathbf{\Lambda}^{1/2}\mathbf{R}'$ be the singular value decomposition of $\mathbf{A}$, where $\mathbf{L}'\mathbf{L} = \mathbf{R}'\mathbf{R} = \mathbf{I}$ and $\mathbf{\Lambda}$ is a diagonal matrix of the eigenvalues of $\mathbf{A}'\mathbf{A}$. Then

$$\begin{aligned}
||\mathbf{u}'\mathbf{A}||^2 &= \mathbf{u}'\mathbf{A}\mathbf{A}'\mathbf{u} \\
&= \mathbf{u}'\mathbf{L}\mathbf{\Lambda}^{1/2}\mathbf{R}'\mathbf{R}\mathbf{\Lambda}^{1/2}\mathbf{L}'\mathbf{u} \\
&= \mathbf{u}'\mathbf{L}\mathbf{\Lambda}\mathbf{L}'\mathbf{u} \\
&\equiv \tilde{\mathbf{u}}'\mathbf{\Lambda}\tilde{\mathbf{u}} \\
&= \sum_{j=1}^{n} \tilde{u}_j^2 \lambda_j,
\end{aligned}$$

where $\tilde{\mathbf{u}} = \mathbf{L}'\mathbf{u}$. We will now identify the distribution of the vector $\{\tilde{u}_1^2, \ldots, \tilde{u}_n^2\}$. Let $\mathbf{B} = \{\mathbf{L}, \mathbf{L}^\perp\}$ be an orthonormal basis for $\mathbb{R}^m$. Since the uniform distribution on the sphere is rotationally invariant, $\mathbf{B}'\mathbf{u}$ is equal in distribution to $\mathbf{u}$, and so $\mathbf{L}'\mathbf{u}$ is equal in distribution to the first n coordinates of $\mathbf{u}$. Recall that a uniformly distributed vector $\mathbf{u}$ can be generated by sampling $z_1, \ldots, z_m$ independently from a standard normal distribution and then dividing each term by $|\sum z_i^2|^{1/2}$. Therefore,

$$\begin{aligned}
\{\tilde{u}_1^2, \ldots, \tilde{u}_n^2\} &\stackrel{d}{=} \frac{\{z_1^2, \ldots, z_n^2\}}{\sum_{j=1}^m z_j^2} \\
&= \left(\frac{\sum_{j=1}^n z_j^2}{\sum_{j=1}^m z_j^2}\right) \left(\frac{\{z_1^2, \ldots, z_n^2\}}{\sum_{j=1}^n z_j^2}\right) \\
&\stackrel{d}{=} \theta\mathbf{q}
\end{aligned}$$

where $\theta \sim \text{beta}(n/2, (m-n)/2)$, $\mathbf{q} \sim \text{Dirichlet}_{\text{n}}(1/2, \ldots, 1/2)$ and θ and $\mathbf{q}$ are independent. Therefore, $||\mathbf{u}'\mathbf{A}||^2 \stackrel{d}{=} \theta\boldsymbol{\lambda}'\mathbf{q}$, where $\boldsymbol{\lambda}$ is the diagonal of $\boldsymbol{\Lambda}$ and are the eigenvalues of $\mathbf{A}'\mathbf{A}$. The required expectation is then

$$E[||\mathbf{u}'\mathbf{A}||^{2l}] = E[\theta^l]E[(\boldsymbol{\lambda}'\mathbf{q})^l]$$

The first expectation is given by $E[\theta^l] = [\Gamma(n/2+l)\Gamma(m/2)]/[\Gamma(m/2+l)\Gamma(n/2)]$. The second expectation is the lth-moment of a Dirichlet average, which results in a type of multiple hypergeometric function denoted as $R_l(\boldsymbol{\lambda}, \frac{1}{2}\mathbf{1})$. This expectation and its generalizations have been studied by Carlson (1977, Chapter 5), Dickey (1983) and others. An algorithm for recursively computing $R_1, \ldots, R_l$ exactly from a generating function is provided in the next section.

To make the result of the calculation a little more intuitive let $\tilde{\boldsymbol{\lambda}} = \boldsymbol{\lambda}/\sum\lambda_j \equiv \boldsymbol{\lambda}/||\mathbf{A}||^2$, and make use of the fact that $E[(\boldsymbol{\lambda}'\mathbf{q})^l] = ||\mathbf{A}||^{2l}E[(\tilde{\boldsymbol{\lambda}}'\mathbf{q})^l]$, Combining the results gives

$$E[e^{\mathbf{u}'\mathbf{A}\mathbf{v}}] = \sum_{l=0}^{\infty} ||\mathbf{A}||^{2l}E[(\tilde{\boldsymbol{\lambda}}'\mathbf{q})^l]\frac{\Gamma(m/2)}{\Gamma(m/2+l)\Gamma(1+l)4^l} \equiv \sum_{l=0}^{\infty} ||\mathbf{A}||^{2l}a_l,$$

and so we see how the expectation is related to the norm of $\mathbf{A}$ via $||\mathbf{A}||^{2l}$ and the variability in relative sizes of the squared singular values via $E[(\tilde{\boldsymbol{\lambda}}'\mathbf{q})^l]$.

To get bounds on a finite-sum approximation to $E[e^{\mathbf{u}'\mathbf{A}\mathbf{v}}]$, note that $\lambda_{\min}^l < E[(\boldsymbol{\lambda}'\mathbf{q})^l] < \lambda_{\max}^l$ so

$$\sum_{l=r+1}^{\infty} \lambda_{\min}^l \frac{\Gamma(m/2)}{\Gamma(m/2+l)\Gamma(1+l)4^l} < \sum_{l=r+1}^{\infty} E[(\boldsymbol{\lambda}'\mathbf{q})^l]\frac{\Gamma(m/2)}{\Gamma(m/2+l)\Gamma(1+l)4^l} < \sum_{l=r+1}^{\infty} \lambda_{\max}^l \frac{\Gamma(m/2)}{\Gamma(m/2+l)\Gamma(1+l)4^l}$$

The outer sums can be computed as

$$\sum_{l=r+1}^{\infty} \lambda^l \frac{\Gamma(m/2)}{\Gamma(m/2+l)\Gamma(1+l)4^l} = \left(\frac{2}{\sqrt{\lambda}}\right)^{m/2-1} I_{m/2-1}(\sqrt{\lambda})\Gamma(m/2) - \sum_{l=1}^{r} \lambda^l \frac{\Gamma(m/2)}{\Gamma(m/2+l)\Gamma(1+l)4^l},$$

and so bounds on $E[e^{\mathbf{u}'\mathbf{A}\mathbf{v}}] - \sum_{l=0}^{r} ||\mathbf{A}||^{2l}a_l$ can be obtained, at least to the same precision with which one can compute the modified Bessel function $I_{m/2-1}(\sqrt{\lambda})$.

C Computing $E[(\boldsymbol{\lambda}'\mathbf{q})^l]$

Let $\mathbf{q} \sim \text{Dirichlet}_n(\alpha_1, \ldots, \alpha_n)$. Carlson (1977, Section 6.6) shows that

$$\prod_{i=1}^{n}(1-t\lambda_i)^{-\alpha_i} = \sum_{l=0}^{\infty} \frac{\Gamma(\boldsymbol{\alpha}'\mathbf{1}+l)}{\Gamma(\boldsymbol{\alpha}'\mathbf{1})\Gamma(l+1)} t^l E[(\boldsymbol{\lambda}'\mathbf{q})^l].$$

Let $c_l = \frac{\Gamma(\boldsymbol{\alpha}'\mathbf{1}+l)}{\Gamma(\boldsymbol{\alpha}'\mathbf{1})\Gamma(l+1)}E[(\boldsymbol{\lambda}'\mathbf{q})^l]$. We now show how to calculate c_{k+1} based on $c_1, \ldots, c_k$. Let $f(t) = \sum_{l=0}^{\infty} c_l t^l$ be the right-hand side of the equation and $g(t) = -\sum_{i=1}^{n} \alpha_i \log(1-t\lambda_i)$ be the log

of the left-hand side. Taking derivatives with respect to t and evaluating at zero we have

$$f^{(l)}(0) = \Gamma(l+1)c_l, \qquad g^{(l)}(0) = \Gamma(l)\sum_{i=1}^{n}\alpha_i\lambda_i^l.$$

Since $f(t) = e^{g(t)}$, we have

$$f^{(k+1)}(0) = \sum_{l=0}^{k}\binom{k}{l}f^{(l)}(0)g^{(k+1-l)}(0).$$

Plugging the values of $f^{(l)}(0)$ into the sum gives

$$c_{k+1} = \sum_{l=0}^{k}\left[c_l\binom{k}{l}\frac{\Gamma(l+1)\Gamma(k+1-l)}{\Gamma(k+2)}\left(\sum_{i=1}^{n}\alpha_i\lambda_i^{k+1-l}\right)\right].$$

Simplifying gives

$$E[(\boldsymbol{\lambda}'\mathbf{q})^{k+1}] = \sum_{l=0}^{k}\left[E[(\boldsymbol{\lambda}'\mathbf{q})^l]\frac{\Gamma(\mathbf{1}'\boldsymbol{\alpha}+l)\Gamma(k+1)}{\Gamma(\mathbf{1}'\boldsymbol{\alpha}+k+1)}\left(\sum_{i=1}^{n}\alpha_i\lambda_i^{k+1-l}\right)\right].$$

C-code with an R-interface to calculate $\{E[(\boldsymbol{\lambda}'\mathbf{q})^l : l = 0, \ldots, k\}$ is available at my website.

References

Carlson, B. C. (1977), *Special functions of applied mathematics*, Academic Press [Harcourt Brace Jovanovich Publishers], New York.

Chikuse, Y. (2003), *Statistics on special manifolds*, vol. 174 of *Lecture Notes in Statistics*, Springer-Verlag, New York.

Dickey, J. M. (1983), "Multiple hypergeometric functions: probabilistic interpretations and statistical uses," *J. Amer. Statist. Assoc.*, 78, 628–637.

Edelman, A. (1988), "Eigenvalues and condition numbers of random matrices," *SIAM J. Matrix Anal. Appl.*, 9, 543–560.

Gabriel, K. R. (1971), "The biplot graphic display of matrices with application to principal component analysis," *Biometrika*, 58, 453–467.

Gabriel, K. R. (1978), "Least squares approximation of matrices by additive and multiplicative models," *J. Roy. Statist. Soc. Ser. B*, 40, 186–196.

Gabriel, K. R. (1998), "Generalised bilinear regression," *Biometrika*, 85, 689–700.

Gower, J. C. and Hand, D. J. (1996), *Biplots*, vol. 54 of *Monographs on Statistics and Applied Probability*, Chapman and Hall Ltd., London.

Green, P. J. (2003), "Trans-dimensional Markov chain Monte Carlo," in *Highly structured stochastic systems*, vol. 27 of *Oxford Statist. Sci. Ser.*, pp. 179–206, Oxford Univ. Press, Oxford, With part A by Simon J. Godsill and part B by Juha Heikkinen.

Harshmanm, R. A., Green, P. E., Wind, Y., and Lundy, M. E. (1982), "A Model for the Analysis of Asymmetric Data in Marketing Research," *Marketing Science*, 1, 205–242.

Hoff, P. D. (2005), "Bilinear mixed-effects models for dyadic data," *J. Amer. Statist. Assoc.*, 100, 286–295.

Householder, A. S. and Young, G. (1938), "Matrix Approximation and Latent Roots," *Amer. Math. Monthly*, 45, 165–171.

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., and Saul, L. K. (1999), "An Introduction to Variational Methods for Graphical Models," *Machine Learning*, 37, 183–233.

Lopes, H. F. and West, M. (2004), "Bayesian model assessment in factor analysis," *Statist. Sinica*, 14, 41–67.

Minka, T. P. (2000), "Automatic choice of dimensionality for PCA," Technical report 514, Media Lab, Massachusetts Institute of Technology.

Mitchell, T. J. and Beauchamp, J. J. (1988), "Bayesian variable selection in linear regression," *J. Amer. Statist. Assoc.*, 83, 1023–1036, With comments by James Berger and C. L. Mallows and with a reply by the authors.

Rajan, J. J. and Rayner, P. J. W. (1997), "Model Order Selection For The Singular Value Decomposition And The Discrete Karhunen-Loeve Transform Using A Bayesian Approach," *Vision, Image and Signal Processing, IEE Proceedings*, 144, 116–123.

Smith, P. J. (1995), "A recursive formulation of the old problem of obtaining moments from cumulants and vice versa," *The American Statistician*, 49, 217–218.

Tierney, L. (1994), "Markov chains for exploring posterior distributions," *Ann. Statist.*, 22, 1701–1762, With discussion and a rejoinder by the author.